

日本国特許庁
JAPAN PATENT OFFICE

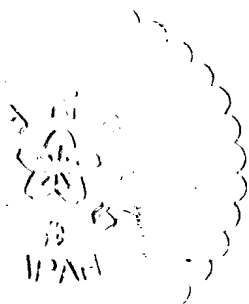
別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2003年 4月21日
Date of Application:

出願番号 特願2003-115185
Application Number:
[ST. 10/C]: [JP2003-115185]

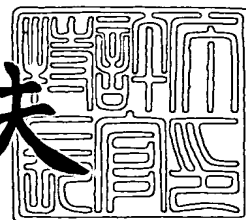
出願人 株式会社日立製作所
Applicant(s):



2003年 9月30日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫



出証番号 出証特2003-3080091

【書類名】 特許願

【整理番号】 K03008231A

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/30

【発明者】

【住所又は居所】 神奈川県川崎市幸区鹿島田 8 9 0 番地 株式会社日立製作所 ビジネスソリューション事業部内

【氏名】 原 憲宏

【発明者】

【住所又は居所】 神奈川県横浜市戸塚区戸塚町 5 0 3 0 番地 株式会社日立製作所 ソフトウェア事業部内

【氏名】 宮崎 光夫

【発明者】

【住所又は居所】 神奈川県横浜市戸塚区戸塚町 5 0 3 0 番地 株式会社日立製作所 ソフトウェア事業部内

【氏名】 河村 信男

【発明者】

【住所又は居所】 神奈川県横浜市戸塚区戸塚町 5 0 3 0 番地 株式会社日立製作所 ソフトウェア事業部内

【氏名】 八高 克志

【発明者】

【住所又は居所】 神奈川県川崎市幸区鹿島田 8 9 0 番地 株式会社日立製作所 ビジネスソリューション事業部内

【氏名】 菅 将孝

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

**【代理人】****【識別番号】** 100075096**【弁理士】****【氏名又は名称】** 作田 康夫**【手数料の表示】****【予納台帳番号】** 013088**【納付金額】** 21,000円**【提出物件の目録】****【物件名】** 明細書 1**【物件名】** 図面 1**【物件名】** 要約書 1**【プルーフの要否】** 要

【書類名】 明細書

【発明の名称】

高可用性を提供するデータベース処理方法

【特許請求の範囲】

【請求項 1】

データベースを複数の分割データベースに分割し、そのおののに対してDBサーバを関連付けてデータ処理を行うデータベース管理システムにおいて、

あらかじめダウン状態の他のDBサーバの処理を代行する関係を示す情報を登録しておくステップと、

あるDBサーバに処理を要求する際に、そのDBサーバがダウンしている場合、該代行する関係を示す情報から代行するサーバを取得し、処理の要求先を変更するステップと、

その要求に代行するための指示を付加するステップと、

その要求を受け取った際に、代行するための指示を判定し、指示が存在する場合、前記ダウンしているサーバの代わりにデータ処理を行うステップとを有することを特徴とするデータベース処理方法。

【請求項 2】

請求項 1 に記載のデータベース処理方法において、
ダウンしているサーバの代わりにデータ処理を行う際に、実行環境をダウンしているサーバの実行環境に切り換えること

を特徴とするデータベース処理方法。

【請求項 3】

請求項 1 に記載のデータベース処理方法において、
ダウンしているサーバの代わりにデータ処理を行う際に、本来のDBサーバと関連付けられたDB格納領域、テーブルあるいはインデクスをアクセスするためのデータベースバッファを利用すること

を特徴とするデータベース処理方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明はデータ処理技術に関し、系切り替え機能を有するデータベース管理システムに適用して有効な技術に関するものである。

【0002】**【従来の技術】**

サービスの停止が重大なビジネスチャンスの喪失に直結するネットビジネスの世界では、24時間365日稼動し続ける堅牢性を備えたシステムが求められる。その中でも特に、障害発生時の影響の局所化と迅速なシステム回復が重要である。従来より、データベース(DB)システムでは、障害発生時の迅速なシステム回復のために、サービス実行用の実行系マシンとは別に待機系マシンと用意し、障害発生時に待機系マシンにサービスの実行を切り替える「系切り替え」という技術が用いられてきた。

【0003】

例えば、文献(Jim Gray and Andreas Reuter, Morgan Kaufmann Publishers, 1993)はHA(High Availability)システム構成によるDB障害対策としてホットスタンバイ無停止運用について開示している。

【0004】

一方、文献Parallel Database Systems :The Future of High Performance Database Systems(David DeWitt and Jim Gray, COMMUNICATIONS OF THE ACM, Vol.35, No.6, 1992, P.85-P.98)には、データベースの処理負荷を複数のプロセッサに分散させ並列に実行するアーキテクチャが開示されている。上記従来技術に記載のShared everything, Shared disk(共用型)アーキテクチャでは、DB処理を実行する全ての計算機が全てのデータをアクセスすることが可能であり、Shared nothing(非共用型)アーキテクチャでは、自計算機に接続されたディスクに格納されたデータのみアクセス可能である。

【0005】

Shared nothing(非共用型)アーキテクチャは、共用型アーキテクチャにくらべ、DB処理を実行する構成単位の間での共有リソースが少なく、スケーラビリティの点で非常に優れている。Shared nothing(非共用型)アーキテクチャにお

いても、高可用性を提供するために系切り換えの技術を用いることが多い。

【0006】

【非特許文献1】

Jim Gray and Andreas Reuter, Morgan Kaufmann Publishers, 1993

【非特許文献2】

Parallel Database Systems :The Future of High Performance Database Systems(David DeWitt and Jim Gray, COMMUNICATIONS OF THE ACM, Vol.35, No.6, 1992, P.85-P.98)

【0007】

【発明が解決しようとする課題】

しかし、系切り替えでは、実行系マシンとは別に待機系マシンとを用意しなければならない、通常サービス実行時には待機系マシンは遊んでいる状態である。また、系切り替えでも「相互待機」という形態で、待機系マシンにも通常のサービス実行を割り当てることもできるが、切り替え時のシステム回復の高速化を図るために、待機系でのシステム立上げをあらかじめ途中まで行っておく(「ウォームスタンバイ」、「ホットスタンバイ」)ことが多く、待機系システムのリソース(プロセスやメモリなど)を余分に用意しなければならない。

【0008】

上記のような通常時未稼動状態である待機専用のリソースを必要とするシステムにおいては通常時にリソースを有効に利用できておらず、システム構築・運用におけるTCO(Total Cost of Ownership)削減の観点で問題である。

【0009】

本発明の目的は上記問題を解決し、前述のような通常時未稼動状態である待機専用のリソースを必要とすることなく、かつ障害発生時に切り替え時間が短い、DBシステム系切り替え制御方法を提供することである。特に、Shared nothingアーキテクチャを用いた並列データベース管理システムにおいて高可用性を提供するデータベース処理方法を提供することを目的とする。

【0010】

【課題を解決するための手段】

前記課題を以下の手段により解決する。

【0011】

Shared nothing(非共用型)アーキテクチャを用いたデータベース管理システムにおいて、あらかじめダウン状態の他のDBサーバの処理を代行する関係を示す情報を登録しておき、ユーザからの問合せを受け付け、あるDBサーバに処理を要求する際に、そのDBサーバがダウンしている場合、該代行する関係を示す情報から代行するサーバを取得し、処理の要求先を変更する。そして、その要求に代行するための指示を付加する。上記要求を受け取ったDBサーバは、代行するための指示を判定し、指示が存在する場合、前記ダウンしているサーバの替わりにデータ処理を行う。

【0012】

前記ダウンしているサーバの替わりにデータ処理を行う際には、替わりを務めるサーバは、実行環境を前記ダウンしているサーバの実行に切り換える。ここで、すでに以前替わりを務めた状態である場合には、実行環境の切り換えは行わない。また、データ処理を行う際のデータベースへのアクセスに使用するデータベースバッファは、本来のDBサーバに関連付けられたDB格納領域へのアクセスに使用しようとしているデータベースバッファを共用する。上記ダウン状態の他のDBサーバの処理を代行する関係を示す情報は、データベース管理システムが内部で自動的に生成してもよい。さらに、代行の優先順位を付加することにより、あるDBサーバがダウンした場合の代行サーバを複数指定することができる。

【0013】

【発明の実施の形態】

Shared nothing(非共用型)アーキテクチャを用いたデータベース管理システムにおいて予備系専用のリソースを用意することなく、障害発生時にDBアクセス処理を即座に再開することが可能な一実施形態のデータベース処理システムについて説明する。

【0014】

まず、本発明の概念を図1を用いて簡単に説明する。

【0015】

本実施例のデータベース管理システムは、処理要求受付サーバ（FES：Front End Server）10およびDBアクセスサーバ（BES：Back End Server）20を具備する。

【0016】

処理要求受付サーバ（FES）10は、ユーザからの問合せ50を受け付け解析し、DBアクセス要求を生成し、DBアクセスサーバへDBアクセス要求を行う。そしてDBアクセスの結果を必要に応じてユーザに返す。DBアクセスサーバ（BES）20は、処理要求受付サーバ10からのDBアクセス要求を受け取り、要求にしたがってDB格納領域上のデータを操作し、必要に応じて結果を処理要求受付サーバに返す。FES10およびBES20とも1プロセスもしくは複数のプロセスによって実現される。

【0017】

本実施例のデータベース管理システムのアーキテクチャは、Shared nothing（非共用型）アーキテクチャであり、本システムが管理するデータベース（例えばテーブル、インデクス）は、さまざまな手法により複数の分割テーブルおよび分割インデクスに分割され、複数のDB格納領域に分散格納される。あるDB格納領域は決まったDBアクセスサーバに対応付けられており、DBアクセスサーバは、そのDBアクセスサーバに対応付けられたDB格納領域内のデータ（例えばテーブルデータ、インデクスデータ）のみをアクセスする。

【0018】

図1の例では、通常、BES1は、DB格納領域1へのアクセス要求のみを、BES2は、DB格納領域2へのアクセス要求のみを処理する。BES1およびBES2は同じDB格納領域をアクセスすることはない。

【0019】

通常時、BES1およびBES2とも稼動状態であり、すべてのリソース（DBアクセスサーバを実現するプロセス、メモリ等）が有効に活用されている。

【0020】

ここで例えば情報処理装置に電源等の障害が発生し、BES1がダウンした場合、ほかの稼動中のサーバ（本例ではBES2）に処理を引き継いでサービスを続

行する。すなわち、B E S 2 が B E S 1 に対する D B アクセス要求を代行する。

【 0 0 2 1 】

具体的には、ユーザ(すなわちアプリケーションプログラム)からの問合せ要求 5 0 を受け取った F E S (1 0) は、問合せ要求を解析し、アクセスするデータが存在する D B 格納領域を決定する。決定された D B 格納領域へのアクセスを担当する D B アクセスサーバがダウン中である場合、障害時の代行情報を基に、処理を代行してくれる代行サーバを決定する (1 3) 。本例では、B E S 1 の代行サーバは B E S 2 であることから、B E S 2 を代行先として決定する。そして、D B アクセス要求 4 0 に代行指示を付加し (1 4) 、D B アクセスサーバ B E S 2 に対してその D B アクセス要求を送信する (1 5) 。代行指示の実装形態は、フラグでも B E S 1 を特定できる識別子でもかまわない。本例ではフラグとして説明する。

【 0 0 2 2 】

F E S (1 . 0) からの D B アクセス要求 4 0 を受け取った D B アクセスサーバ B E S 2 は、D B アクセス要求が B E S 2 が担当している D B 格納領域 2 への要求であるか、別の D B 格納領域へのアクセス要求であるかを判定する (2 0) 。判定には、D B アクセス要求に付加された代行指示を用いる。代行指示のフラグが O N の場合、障害時の代行情報を基に本来 B E S 1 が担当している D B 格納領域 1 へのアクセス要求であることを認識する。そして、D B アクセスサーバ B E S 1 の実行環境に切り換える (2 4) 。例えば、D B アクセスサーバ B E S 2 を実現しているプロセスの環境変数やメモリ情報を B E S 1 のものに変更する。

【 0 0 2 3 】

切り換えた実行環境の下、D B アクセス処理を実行し (2 5) 、D B 格納領域 1 をアクセスし要求されたデータ操作を行う。本例では、データベース 6 0 内の D B 格納領域 1 (6 1) に格納されたテーブルデータ (6 2) の " 1 2 " を実行結果として、F E S に送信する。送信されたテーブルデータは問合せ結果として F E S (1 0) によってユーザに返却される。

【 0 0 2 4 】

ここでの障害時の代行情報は、D B M S の管理者などのユーザによりあらかじ

めDBMSに登録されているものとするが、DBMSが内部に自動生成することにより管理者の負担を低減することができる。

【0025】

障害が発生した際、ダウン中のDBアクセスサーバBES1の代わりにDB格納領域1へのアクセス要求を代行して処理してくれるDBアクセスサーバBES2をここでは、「代行サーバ」と呼ぶ。代行サーバとして動作する際に、あらたなプロセスが生成されるわけではなく、もともとのBESのプロセスのまま、代行サーバとして処理を行うため、無駄なプロセス生成もない。

【0026】

以上のように、別のサーバを代行サーバとして登録もしくは決定しておき、障害発生時には、FESでBES未稼動状態を検知し、代行指示を用いて稼動中のサーバに処理の代行をさせることにより、予備系専用のリソースなしで、障害発生時に即座にDBアクセス処理を再開することが可能である。

【0027】

ここでは、FESとBESとを別々の情報処理装置上に配置したが、同一の情報処理装置上に配置することによりハードウェア資源の有効活用が可能である。また、本実施例で示したFESの機能およびBESの機能を1つのDBサーバとして実装することにより、データベース管理システムの管理者は、FESとBESとを意識して管理する必要がなくなり、管理コストを低減することができる。

【0028】

次に図2に本実施形態のデータベース管理システムの概略構成を示す。

【0029】

ユーザが作成したアプリケーションプログラム6と、問い合わせやリソース管理などのデータベースシステム全体の管理を行うデータベース管理システム2がある。上記のデータベース管理システム2は、処理要求受付サーバ(FES)10、DBアクセスサーバ(BES)20を具備する。また、データベース管理システム2はデータベースバッファ230を具備し、データベースアクセス対象となるデータを永続的にあるいは一時的に格納するデータベース3、そして障害時の代行情報30を有する。

【0030】

処理要求受付サーバ（FES）10は、アプリケーションプログラム6から投入される問合せを受け付け解析し、DBアクセス要求を生成し、DBアクセスサーバへDBアクセス要求を行う。そしてDBアクセスの結果を必要に応じてアプリケーションプログラム6に返す。DBアクセスサーバ（BES）20は、処理要求受付サーバ10からのDBアクセス要求を受け取り、要求にしたがって外部記憶装置上に記憶されるデータベース3をデータベースバッファ230を通じてアクセスする。先に図1において説明した代行サーバのDBアクセス処理では、もともとのBESが使用しているものと同じデータベースバッファを利用する。すなわち、もともとのBESと代行サーバ（BES）とではデータベースバッファを共用する。

【0031】

上記データベース管理システム2はネットワークなどを介して他のシステムと接続されている。また、処理要求受け付けサーバ（FES）10およびDBアクセスサーバ（BES）20は必ずしも同一の情報処理装置上に配置される必要はない。それぞれ別々の情報処理装置上に配置され、ネットワーク等を介して1つのデータベース管理システムとして機能すればよい。また、1つのデータベース管理システムは複数のFESを配置することにより多量のユーザからの要求の負荷を分散させることができる。また、複数のBESを有することによりデータ処理の並列度が高まり、大規模なデータベースに対するデータ処理も高速に実現することができる。

【0032】

上記処理要求受付サーバ10は、問合せの構文解析・意味解析を行い、適切な処理手順を決定し、その処理手順に対応したコードを生成し、DBアクセスサーバ20に対しDBアクセス要求を行う処理要求制御部211を具備している。また、上記処理要求受付サーバ10は、処理要求制御部211においてDBアクセス要求を行う際に、要求先DBアクセスサーバの稼動状況を判定し、必要に応じてデータ処理要求先を変更し（213）、変更先に代行を支持する（214）代行制御部F（212）を具備している。

【0033】

上記DBアクセスサーバ20は、処理要求受付サーバ10から受け取ったDBアクセス要求（生成したコード）にしたがってデータベース3上のデータのアクセス制御等を行うデータ処理制御部221を具備している。また、上記DBアクセスサーバ20は、データ処理制御部221においてDBアクセス要求を受け取った際に、代行処理要求であるかを判定し、代行サーバとしての実行環境への切り替え（223）等の制御を行う代行処理制御部B（222）を具備している。また、代行処理制御部Bは、代行元のサーバに障害が発生した際に、関連するDB格納領域ほかの回復を行い、実行中だった処理の更新結果を取り消すトランザクション回復を行う代行部回復制御機能を有する。

【0034】

図3は本実施形態のコンピュータシステムのハードウェア構成の一例を示す図である。

この例のコンピュータシステムは、情報処理装置3000、3100および3200を含む。

【0035】

情報処理装置3000は、CPU3002、主記憶装置3001、通信制御装置3003、I/O制御装置3004及び端末3006により構成される。主記憶装置3001上には、アプリケーションプログラム3008が置かれ、CPU3002を用いて稼動している。アプリケーションプログラム3008がDBMS2の処理要求受付サーバ10にユーザ問合せ50を行うと、情報処理装置3000の通信制御装置3003と情報処理装置3100の通信制御装置3003によって、ネットワーク3007を経由して処理要求受付サーバ10に問合せ要求が送られる。

【0036】

情報処理装置3100は、CPU3002、主記憶装置3001、通信制御装置3003、I/O制御装置3004、磁気ディスク装置等の外部記憶装置3005及び端末3006により構成される。情報処理装置3100の主記憶装置3001上には、図2を用いて先に説明した処理要求受付サーバ10を有するデー

データベース管理システム 2 が置かれ、CPU 3002 を用いて稼動している。外部記憶装置 3005 上にはデータベース管理システム 3 が管理するデータベース 3 が格納される。また、データベース管理システム 2 を実現するプログラム 3100 も外部記憶装置 3003 上に格納される。処理要求受付サーバ 10 は、I/O 制御装置 3004 により外部記憶装置 3005 からデータの読み出し/書き込みを行い、通信制御装置 3003 によりネットワーク 3007 で接続された他の情報処理装置とデータの送受信を行う。

【0037】

情報処理装置 3200 は、CPU 3002、主記憶装置 3001、通信制御装置 3003、I/O 制御装置 3004、磁気ディスク装置等の外部記憶装置 3005 及び端末 3006 により構成される。情報処理装置 3200 の主記憶装置 3001 上には、図 2 を用いて先に説明した DB アクセスサーバ 20 を有するデータベース管理システム 2 が置かれ、CPU 3002 を用いて稼動している。外部記憶装置 3005 上にはデータベース管理システム 3 が管理するデータベース 3 が格納される。また、データベース管理システム 2 を実現するプログラム 3100 も外部記憶装置 3003 上に格納される。DB アクセスサーバ 20 は、I/O 制御装置 3004 により外部記憶装置 3005 からデータの読み出し/書き込みを行い、通信制御装置 3003 によりネットワーク 3007 で接続された他の情報処理装置とデータの送受信を行う。

【0038】

ここで、2つの情報処理装置 3200 にまずそれぞれ関連付けられたデータベース 3 が格納された外部記憶装置 3005 は、共用ディスクであり、他方の情報処理装置からのアクセスも可能である。データベース管理システム 2 の稼動および系切り替え操作を制御するクラスタウェアなどにより、上記共用ディスクのアクセス制御は行われる。

【0039】

図 4 は本実施形態の処理要求制御部および代行制御部 F の処理手順の一部を示すフローチャートである。

【0040】

まず、ステップ401において、ユーザからの問合せを受け取り、ステップ402において、ユーザからの問合せを解析して得られた情報から、問合せを実現するためのDB格納領域へのアクセスを担当しているDBアクセスサーバを確定する。次にステップ403において、確定したDBアクセスサーバが稼動中であるかどうかを判定する。そのDBアクセスサーバが稼動中である場合、ステップ407に進み、確定したDBアクセスサーバに対して処理要求を送信する。

【0041】

また、ステップ403において、確定したDBアクセスサーバがダウンしていて未稼動状態である場合、ステップ404に進み、当該DBアクセスサーバに関連する障害時の代行情報を取得する。そして、取得した障害時の代行情報30から代行サーバを代行先に決定する（ステップ405）。次に、処理要求へ代行指示を付加、すなわち代行指示のフラグをONにする（ステップ406）。ステップ407に進み、代行先に決定したDBアクセスサーバに対して処理要求を送信する。

【0042】

図5は本実施形態のデータ処理制御部および代行制御部Bの処理手順の一部を示すフローチャートである。

【0043】

DBアクセス処理は、図4のフローチャートで説明した処理要求の送信に引き続き行われる。

【0044】

まず、ステップ501において、FESから実行要求を受け取り、受け取った実行要求が代行指示かどうかの判定を行う（ステップ502）。代行指示のフラグがONの場合、ステップ503に進み、自DBアクセスサーバが代行サーバに指定されている障害時の代行情報を取得する（ステップ503）。そして、取得した障害時の代行情報30から代行指示のあったサーバ名を参照し、そのサーバに関する実行環境へ切り換える（ステップ504）。切り換えた実行環境の下でDBアクセス処理を実行し、ダウン中のDBアクセスサーバに代わってそのDB格納領域をアクセスしデータ操作を行う（ステップ505）。

【0045】

また、ステップ502の判定において代行指示のフラグがOFFの場合には、実行環境の切り換えは行わず、もともとのDBアクセスサーバとして自分の担当するDB格納領域をアクセスし、データ操作を行う（ステップ505）。

【0046】

ステップ504において、稼動中の実行環境のチェックを行い、処理要求を実行するためのBESと、現稼動中のBESとが同じ場合には、実行環境を切り替えないような制御を行うことも考えられる。また、ステップ505においてDBアクセス処理実行後に、もともとのBESの実行環境に戻すことも考えられる。それらは、処理要求を制御するバランサもしくはスケジューラによって最適な制御を行うことが可能である。

【0047】

図6は本実施形態の代行情報の一例を示す図である。

図6（a）の例では、障害時の代行情報30は、DBアクセスサーバのサーバ名と、そのサーバがダウンした場合に、対応するDB格納領域へのアクセスを替わりに行う代行サーバのサーバ名より構成される。代行情報600は、DBアクセスサーバBES1が障害によりダウンした場合に、BES2が代行サーバとして処理を代行することを示している。

【0048】

また、図7および図8は、図6の（b）、（c）、（d）および（e）にそれぞれ対応した代行サーバ構成を示す図である。

【0049】

図6（b）の代行情報の例は、図7（a）の代行サーバ構成に対応している。図6（b）の例では、情報処理装置701上のDBアクセスサーバBES3（705）と、情報処理装置702上のDBアクセスサーバBES4（706）は、図6（c）の代行情報601および602によって、それぞれ相互に代行サーバとして指示されている。すなわち相互代行の構成をとっている。具体的には、BES3がダウンした場合には、BES3の処理をBES4が代行する（代行情報601）。BES4がダウンした場合には、BES4の処理をBES3が代行す

る（代行情報 602）。

【0050】

さらに、情報処理装置 703 上の BES 5 および BES 6 と、情報処理装置 704 上の BES 7 および BES 8 も、2つの情報処理装置間でそれぞれの BES が相互代行構成をとっている。具体的には、BES 5 がダウンした場合には、BES 7 が処理を代行し、BES 6 がダウンした場合には、BES 8 が処理を代行する。また、BES 7 がダウンした場合には、BES 5 が処理を代行し、BES 8 がダウンした場合には、BES 6 が処理を代行する。

【0051】

図 7（b）の表記において、例えば 705 の表記 “（BES 4*）” の ‘（’ および ‘）’ は、その DB サーバが “BES 4” としては稼動していない、すなわち DB 格納領域 RD 4 をアクセスする DB 処理は行っていない（BES 3 として稼動している）ことを示す。また、表記 ‘*’ は、DB サーバ 705 が表記 ‘*’ を付加した DB サーバ（BES 4）の代行を行うよう指定されていることを示す。

【0052】

図 7（a）におけるサーバの稼動状況は、情報処理装置 701 および 703 に障害が発生し BES 3、BES 5 および BES 6 がダウンした状態である。DB 格納領域 RD 3 へのアクセスを伴うデータ操作要求は、代行指示を付加され DB アクセスサーバ 706 に送信される。要求を受け取った DB アクセスサーバ 706 のプロセスは、BES 3 として処理を代行し RD 3 をアクセスする。

【0053】

RD 5 へのアクセスを伴うデータ操作要求も同様に、代行指示を付加され DB アクセスサーバ 707 に送信される。要求を受け取った DB アクセスサーバ 707 のプロセスは、BES 5 として処理を代行し RD 5 をアクセスする。

【0054】

図 6（c）の代行情報の例は、図 7（b）の代行サーバ構成に対応している。図 6（c）の例では、情報処理装置 712 上の DB アクセスサーバ BES 9（715）、情報処理装置 713 上の DB アクセスサーバ BES 10（716）およ

びDBサクセスサーバBES11（717）は、図6（c）の代行情報607、608および609によって、それぞれ片方向に代行サーバとして指示されている。すなわち片方向代行のリング構成をとっている。具体的には、BES9がダウンした場合には、BES9の処理をBES10が代行する（代行情報607）。BES10がダウンした場合には、BES10の処理をBES11が代行する（代行情報608）。そして、BES11がダウンした場合には、BES11の処理をBES9が代行する（代行情報609）。

【0055】

図7（b）におけるサーバの稼動状況は、情報処理装置712に障害が発生しBES9がダウンした状態である。DB格納領域RD9へのアクセスを伴うデータ操作要求は、代行指示を付加されDBアクセスサーバ716に送信される。要求を受け取ったDBアクセスサーバ716のプロセスは、BES9として処理を代行しRD9をアクセスする。

【0056】

図6（d）の代行情報の例は、図8（a）の代行サーバ構成に対応している。図6（d）の例では、情報処理装置801上のDBアクセスサーバBES12（804）、情報処理装置802上のDBサクセスサーバBES12（805）およびDBサクセスサーバBES14（806）は、図6（d）の代行情報610、611および612によって、それぞれ代行サーバが指示されている。特に、代行情報611および612によって、BES12はBES13およびBES14の2つのサーバの代行サーバとして指定されている。すなわち、N：1の代行構成になっている。

【0057】

図8（a）におけるサーバの稼動状況は、情報処理装置803に障害が発生しBES14がダウンした状態である。DB格納領域RD14へのアクセスを伴うデータ操作要求は、代行指示を付加されDBアクセスサーバ801に送信される。要求を受け取ったDBアクセスサーバ801のプロセスは、BES14として処理を代行しRD14をアクセスする。

【0058】

図6 (e) の代行情報の例は、図8 (b) の代行サーバ構成に対応している。図6 (e) の例では、情報処理装置808上のDBアクセスサーバBES12 (804) は、図6 (d) の代行情報613、614および615によって、BES13 (805)、BES14 (806) およびBES15 (807) を代行サーバとして複数指示している。BES12がダウンしている場合、BES12の処理を代行するサーバは、代行情報の代行優先順の値によって決定する。すなわちN段代行構成になっている。

【0059】

まずは、代行情報613の指定により、BES16が代行サーバの候補となる。ここでもしBES16もダウンしている場合は、代行優先順に従い代行情報614のBES17が代行サーバの候補となる。さらにBES17もダウンしている場合には同様にBES18が代行サーバとなる。

【0060】

図8 (b) におけるサーバの稼働状況は、情報処理装置808に障害が発生しBES15がダウンした状態である。さらに情報処理装置809の障害によってBES16もダウンしているため、DB格納領域RD15へのアクセスを伴うデータ操作要求は、代行情報614に従い代行指示を付加されDBアクセスサーバ814に送信される。要求を受け取ったDBアクセスサーバ814のプロセスは、BES15として処理を代行しRD15をアクセスする。

【0061】

以上図4および図5で示したフローチャートの処理は、図3で例として示したコンピュータシステムにおけるプログラムとして実行される。しかし、そのプログラムは図3の例の様にコンピュータシステムに物理的に直接接続される外部記憶装置に格納されるものと限定はしない。ハードディスク装置、フレキシブルディスク装置等のコンピュータで読み書きできる記憶媒体に格納することができる。また、ネットワークを介して図3のコンピュータシステムを構成する情報処理装置とは別の情報処理装置に接続される外部記憶装置に格納することもできる。

【0062】

【発明の効果】

本発明によれば、Shared nothing(非共用型)アーキテクチャを用いたデータベース管理システムにおいて、通常時に稼動していない待機専用のリソースを有することなく、障害発生時にDB処理サービスをすぐに再開することが可能なデータベース処理方法およびシステムを提供することができる。

【図面の簡単な説明】

【図 1】

本発明の概念図である。

【図 2】

本実施形態のデータベース処理システムの機能ブロックを示す図である。

【図 3】

本実施形態のコンピュータシステムのハードウェア構成の一例を示す図である。

。

【図 4】

本実施形態の処理要求制御部および代行制御部 F の処理手順の一部を示すフローチャートである。

【図 5】

本実施形態のデータ処理制御部および代行制御部 B の処理手順の一部を示すフローチャートである。

【図 6】

本実施形態の代行情報の一例を示す図である。

【図 7】

本実施形態の代行サーバ構成の一例を示す図である。

【図 8】

本実施形態の代行サーバ構成の一例を示す図である。

【符号の説明】

- 1 0 …処理要求受付サーバ
- 2 0 …DB アクセスサーバ
- 3 0 …障害時の代行情報
- 4 0 …DB アクセス要求

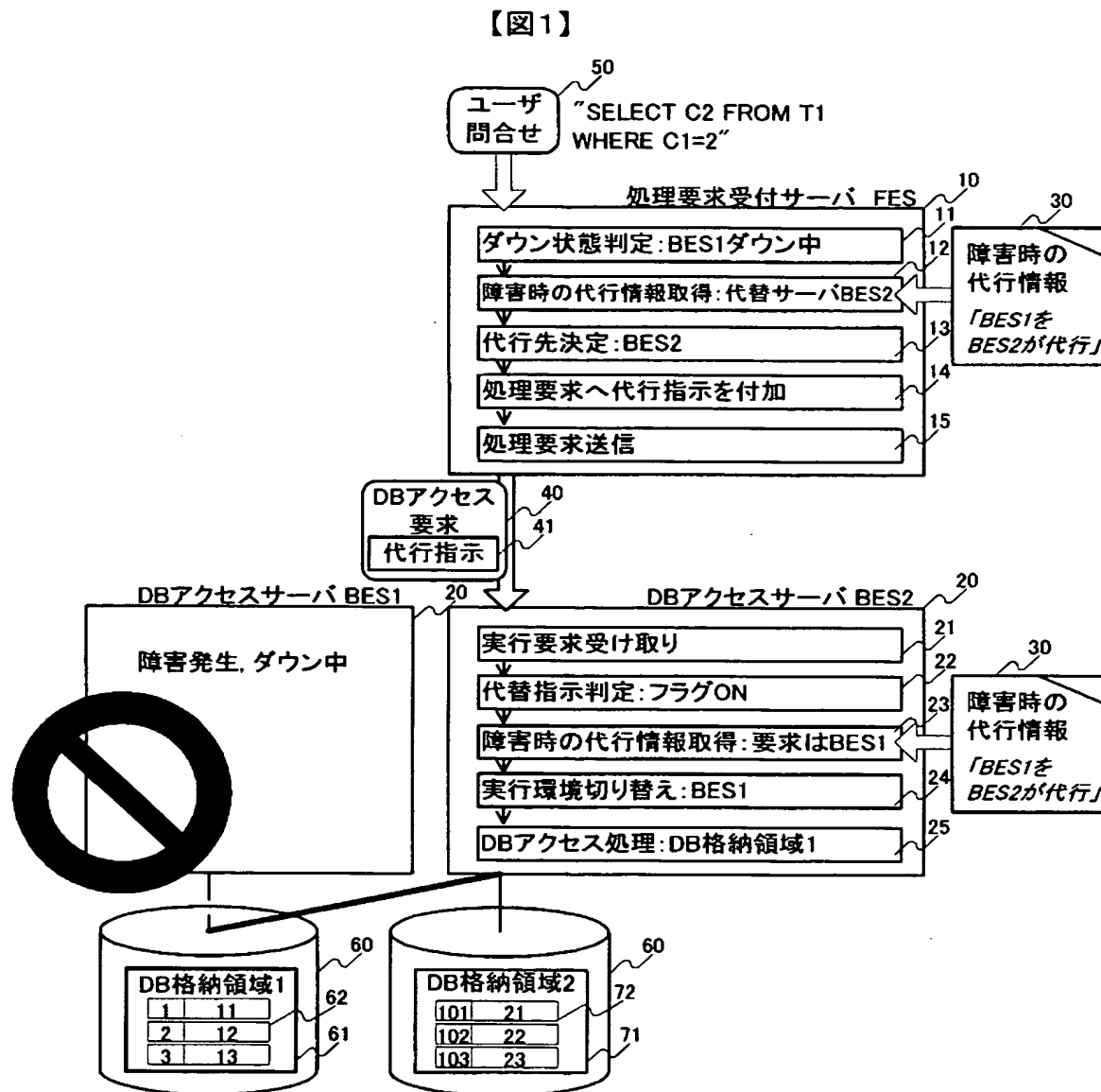
4 1 …代行指示

5 0 …ユーザ問合せ

6 0 …データベース。

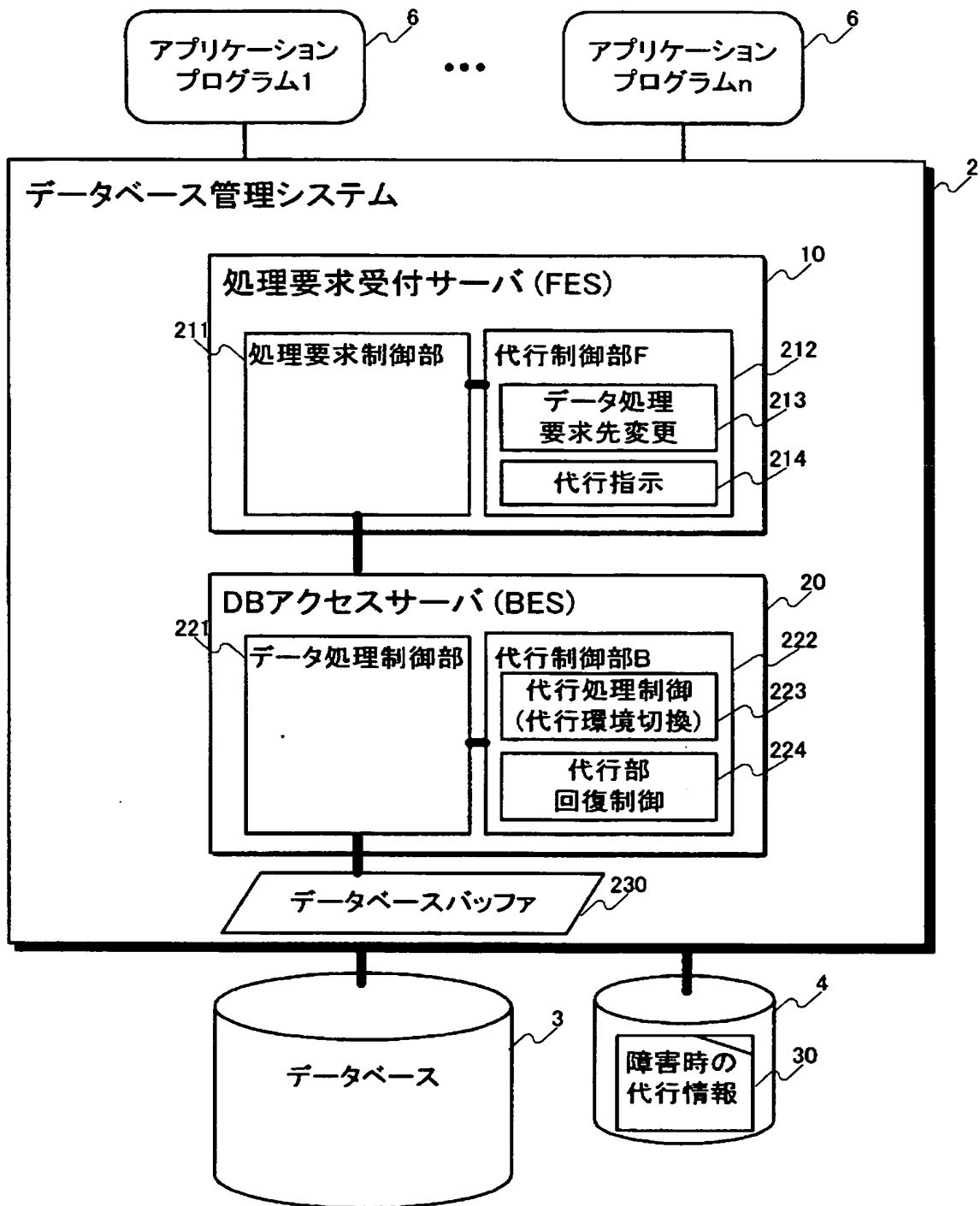
【書類名】 図面

【図1】



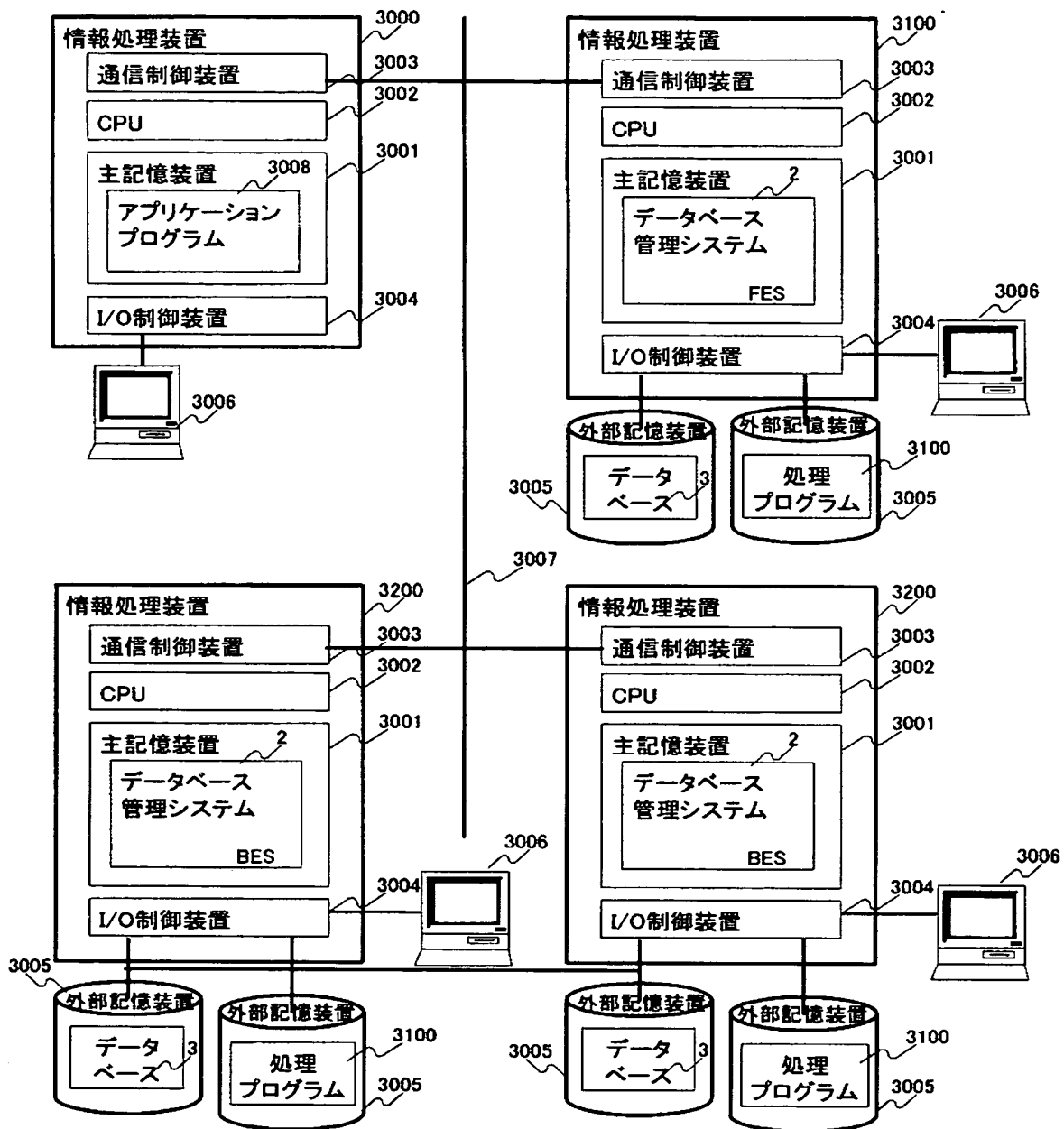
【図2】

【図2】



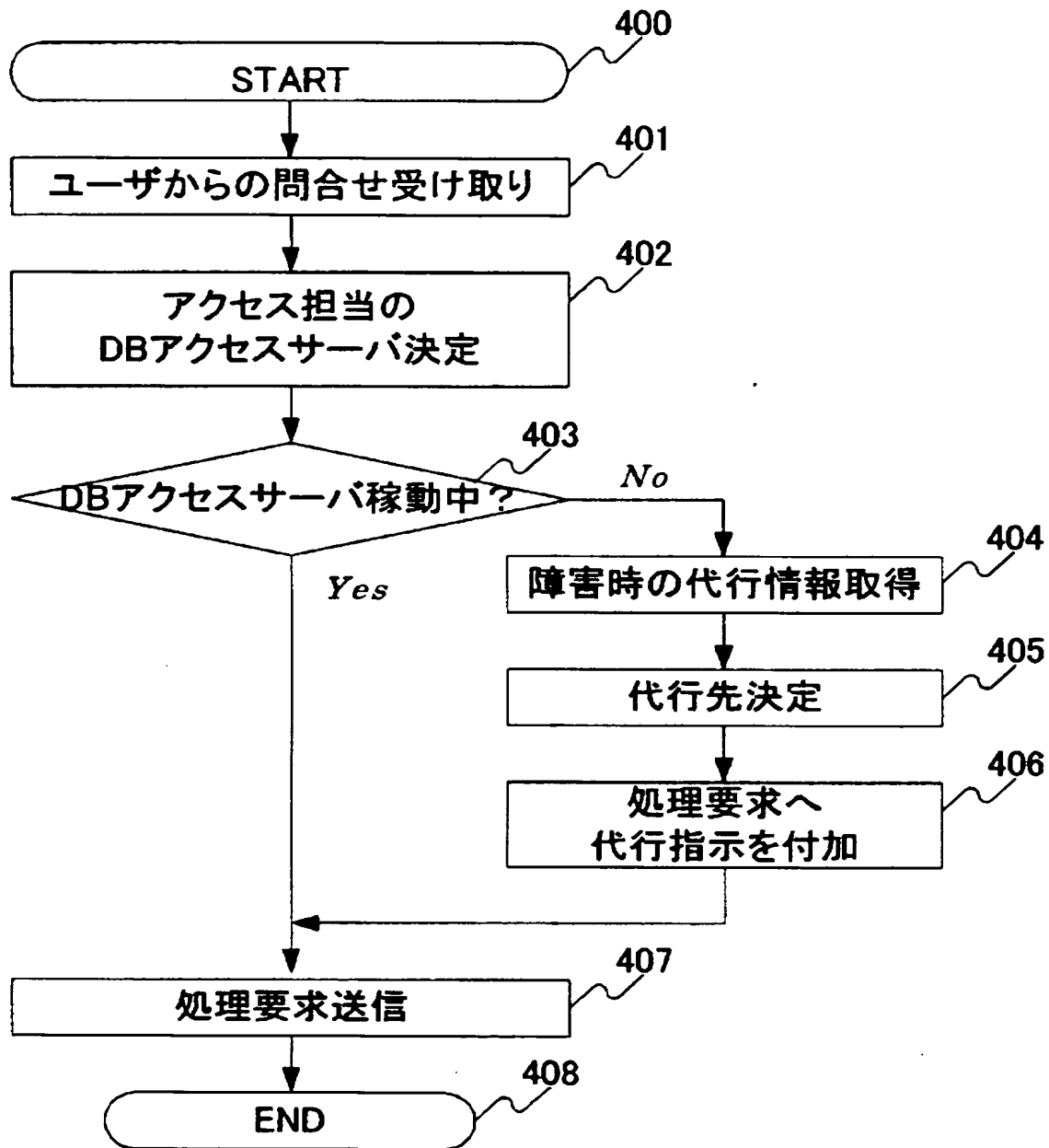
【図 3】

【図3】



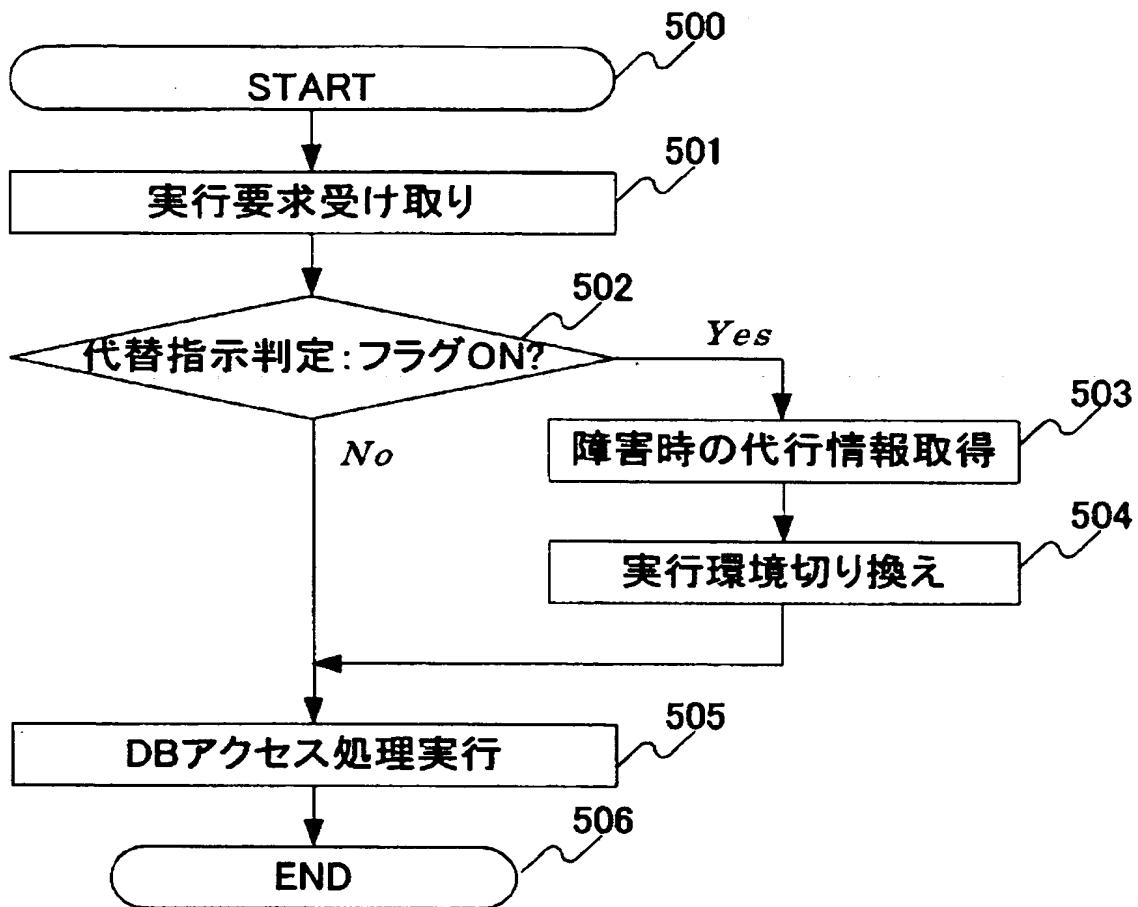
【図 4】

【図 4】



【図5】

【図5】



【図 6】

【図 6】

(a)	サーバ	代行サーバ	600
	BES1	BES2	

(b)	サーバ	代行サーバ	601
	BES3	BES4	602
	BES4	BES3	603
	BES5	BES7	604
	BES6	BES8	605
	BES7	BES5	606
	BES8	BES6	

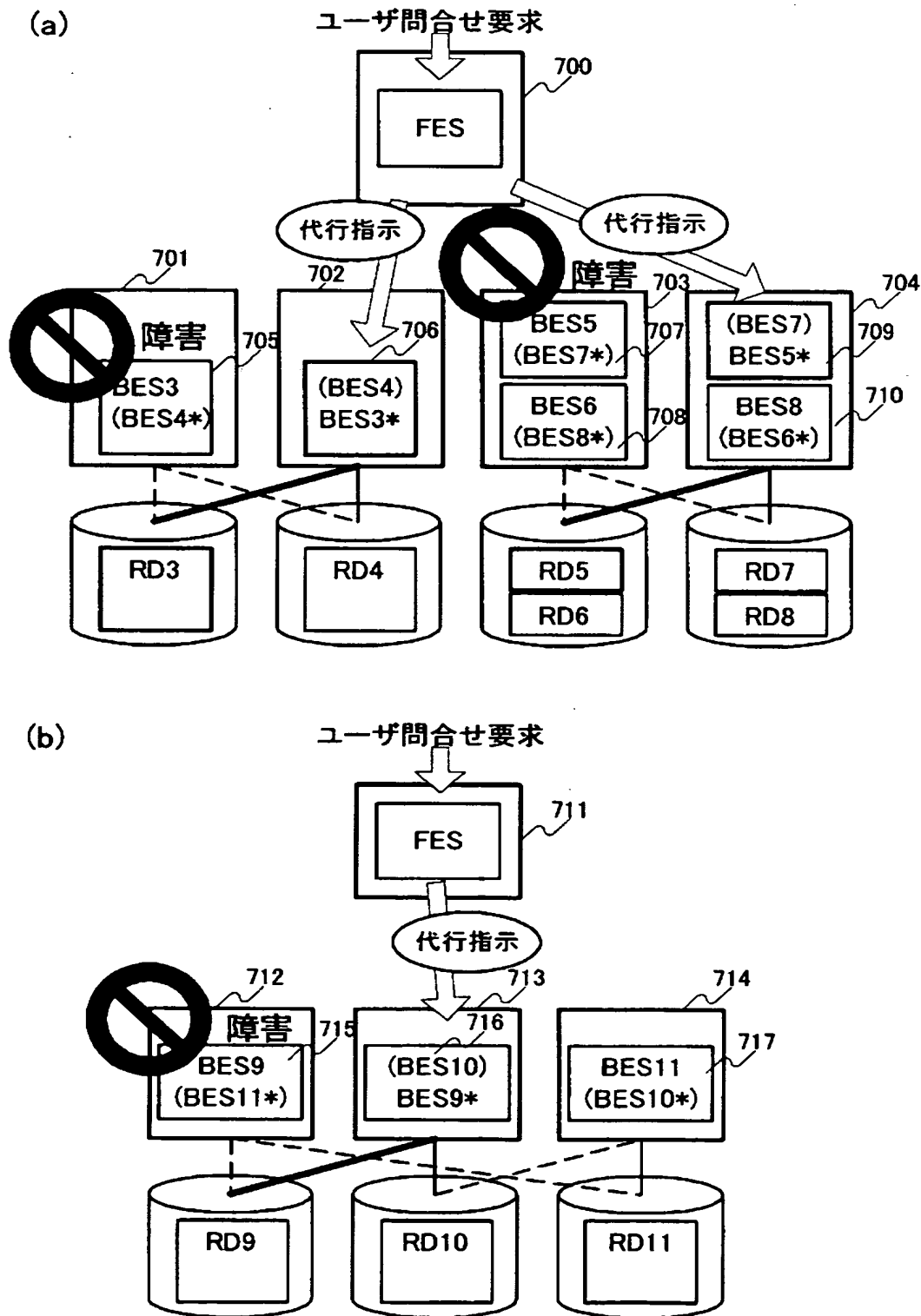
(c)	サーバ	代行サーバ	607
	BES9	BES10	608
	BES10	BES11	609
	BES11	BES9	

(d)	サーバ	代行サーバ	610
	BES12	BES13	611
	BES13	BES12	612
	BES14	BES12	

(e)	サーバ	代行サーバ	代行優先順	613
	BES15	BES16	1	614
	BES15	BES17	2	615
	BES15	BES18	3	
	

【図 7】

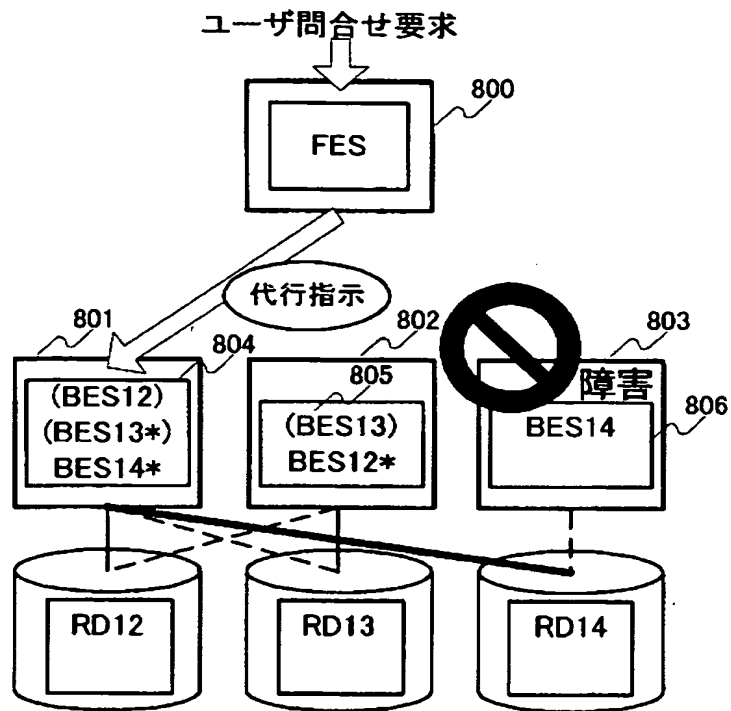
【図 7】



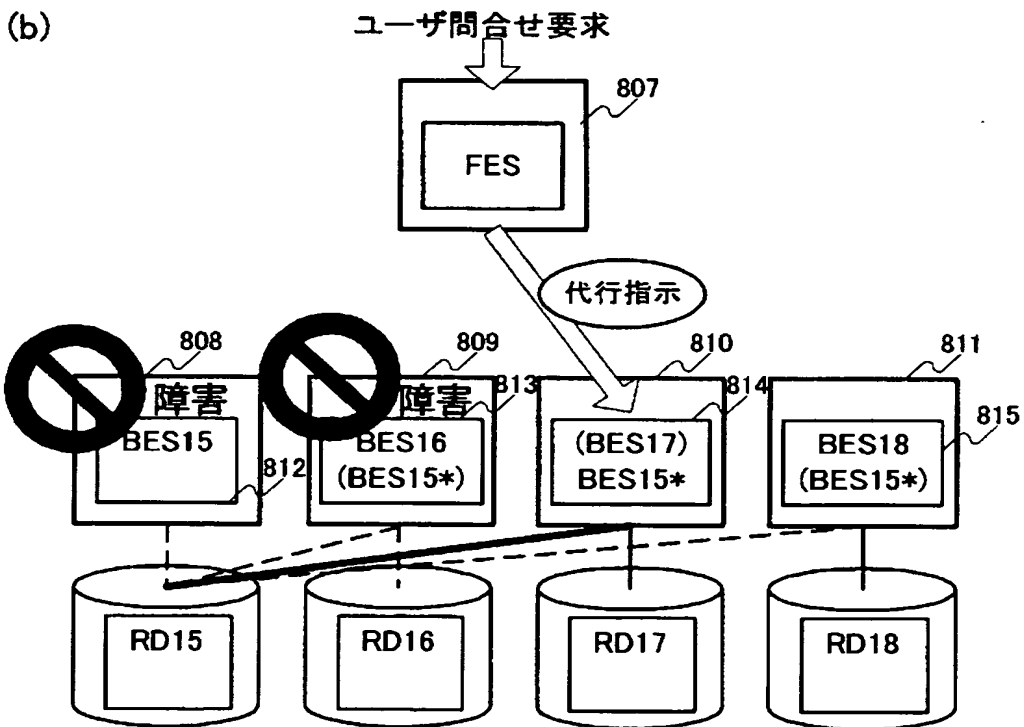
【図 8】

【図 8】

(a)



(b)



【書類名】 要約書**【要約】****【課題】**

データベース管理システム、特に、Shared nothingアーキテクチャを用いた並列データベース管理システムにおいて、通常時未稼動状態である待機専用のリソースを必要としない系切り替え機能を提供する。

【解決手段】

ダウン状態の他のDBサーバの処理を代行する関係を示す情報を登録しておき、ユーザからの問合せを受け付け、あるDBサーバに処理を要求する際に、そのDBサーバがダウンしている場合、該代行する関係を示す情報から代行するサーバを取得し、処理の要求先を変更する。その要求に代行するための指示を付加する。上記要求を受け取ったDBサーバは、代行するための指示を判定し、指示が存在する場合、前記ダウンしているサーバの代わりにデータ処理を行う。

これにより、通常時に稼動していない待機専用のリソースを有することなく、障害発生時にDB処理サービスをすぐに再開することが可能なデータベース処理方法およびシステムを提供することができる。

【選択図】 図 1

認定・付加情報

特許出願の番号	特願 2 0 0 3 - 1 1 5 1 8 5
受付番号	5 0 3 0 0 6 5 1 9 6 4
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 4 月 2 2 日

< 認定情報・付加情報 >

【提出日】	平成15年 4月21日
-------	-------------

次頁無

特願 2 0 0 3 - 1 1 5 1 8 5

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所